From Contrast to Consistency: Rethinking Event-based Continuous-Time Optical Flow Estimation

Supplementary Material

The supplementary material contains the following content. In Sec. A, we provide the motivation and derivation for the bidirectional temporal modeling, along with its enhanced robustness in occluded regions. Sec. B details the implementation of the entire framework. In Sec. C, we present a computational efficiency analysis and additional ablation studies. Sec. D includes more visual results to further validate the effectiveness of our approach. Finally, in Sec. E, we discuss the limitations of our method and outline potential future research directions.

A. Theoretical Motivation for Bidirectional Temporal Modeling

In our main paper, the Bidirectional Refinement Update (BRU) module is introduced as a key component for robust and accurate continuous-time optical flow estimation. This design is theoretically motivated by the superior properties of central-difference approximations compared to unidirectional (backward-difference) schemes, particularly in handling dynamic scenes with non-zero acceleration and occlusions.

A.1. Physical Model of Event Dynamics

Events are not random noise but structured spatio-temporal projections induced by the relative motion between the scene and the camera. Consider a physical point moving along a continuous trajectory $\mathbf{P}(t)$ on the image plane. The instantaneous optical flow $\mathbf{u}(t) \triangleq d\mathbf{P}/dt$ governs the event generation process through the brightness change constraint:

$$|\nabla L \cdot \mathbf{u}(t)\, \Delta t_e| \approx |\Delta L| \geq C, \tag{1}$$

where ∇L denotes the spatial gradient of log-brightness, Δt_e represents the temporal interval since the last event at that pixel, and C is the contrast threshold determining event generation.

This relationship creates a strong coupling between the continuous motion field and the discrete event stream. Therefore, a valid flow estimate must not only maximize contrast but also ensure that the events, when back-projected along $\mathbf{P}(t)$, reconstruct the stable underlying photometric structure. Motivated by this, we propose **Spatio-temporal Structural Consistency (STSC)**, which leverages this physical prior to regularize the motion field, ensuring that the recovered trajectory preserves both local structural sharpness and temporal continuity across the motion manifold.

A.2. Analysis of First-Order Bias in Unidirectional Models

Existing event-based optical flow methods [2–7] predominantly rely on unidirectional recurrent architectures (e.g., forward-pass GRUs). When estimating the instantaneous velocity $\mathbf{u}(t)$ at time t, these models are constrained to use only historical state information $\mathbf{P}(t-\Delta t)$. Mathematically, this is equivalent to a **backward-difference** numerical approximation:

$$\mathbf{u}_f(t) \approx \frac{\mathbf{P}(t) - \mathbf{P}(t-\Delta t)}{\Delta t}. \tag{2}$$

To quantify the estimation error, we perform a Taylor series expansion of $\mathbf{P}(t-\Delta t)$ around the current time t:

$$\mathbf{P}(t-\Delta t) = \mathbf{P}(t) - \mathbf{u}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 - \mathcal{O}(\Delta t^3), \tag{3}$$

where $\mathbf{u}(t) = \dot{\mathbf{P}}(t)$ and acceleration $\mathbf{a}(t) = \ddot{\mathbf{P}}(t)$. Rearranging terms to solve for the backward estimator $\mathbf{u}_f(t)$ reveals a systematic error term:

$$\mathbf{u}_f(t) \approx \mathbf{u}(t) - \underbrace{\frac{1}{2}\mathbf{a}(t)\Delta t}_{\text{First-Order Bias}} + \mathcal{O}(\Delta t^2). \tag{4}$$

Implication: Eq. 4 demonstrates that unidirectional models inherently suffer from a first-order bias proportional to acceleration. In dynamic scenes ($\mathbf{a}(t) \neq 0$), this bias introduces a systematic lag, causing the estimated flow to deviate from the true physical motion. Due to the high temporal resolution of event cameras, first-order errors can easily accumulate over time, leading to a compounding effect that significantly distorts the flow estimation. Consequently, the event generation constraint (Eq. 1) is violated:

$$|\nabla L \cdot \mathbf{u}_f(t)\, \Delta t| \approx |\Delta L(t)| - \frac{1}{2}|\nabla L \cdot \mathbf{a}(t)\Delta t^2|. \tag{5}$$

This systematic misalignment hinders the enforcement of Spatio-temporal Structural Consistency (STSC), as the warped event volume (VWE) will exhibit structural blurring due to the inaccurate velocity field.

A.3. Unbiased Estimation via Bidirectional Modeling

To eliminate the acceleration-induced bias, our Bidirectional Refinement Update (BRU) module integrates both

past and future temporal contexts. This formulation effectively implements a **central-difference** estimator:

$$\mathbf{u}_c(t) \approx \frac{\mathbf{P}(t+\Delta t) - \mathbf{P}(t-\Delta t)}{2\Delta t}. \tag{6}$$

We analyze this estimator by expanding both the future term $\mathbf{P}(t+\Delta t)$ and past term $\mathbf{P}(t-\Delta t)$ via Taylor series:

$$\mathbf{P}(t+\Delta t) = \mathbf{P}(t) + \mathbf{u}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 + \mathcal{O}(\Delta t^3), \tag{7}$$

$$\mathbf{P}(t-\Delta t) = \mathbf{P}(t) - \mathbf{u}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 - \mathcal{O}(\Delta t^3). \tag{8}$$

Subtracting the two expansions leads to the cancellation of all even-order derivatives, including the acceleration term $\mathbf{a}(t)$:

$$\mathbf{u}_c(t) \approx \mathbf{u}(t) + \mathcal{O}(\Delta t^2). \tag{9}$$

Conclusion: The bidirectional modeling achieves second-order accuracy and is physically unbiased with respect to acceleration. By providing an unbiased velocity estimate $\mathbf{u}_c(t)$, the BRU module ensures that the warped event volumes used in $\mathcal{L}_{\text{LSC}}$ and $\mathcal{L}_{\text{TC}}$ accurately reflect the underlying physical motion.

A.4. Practical Benefits: Robustness in Degraded Regions

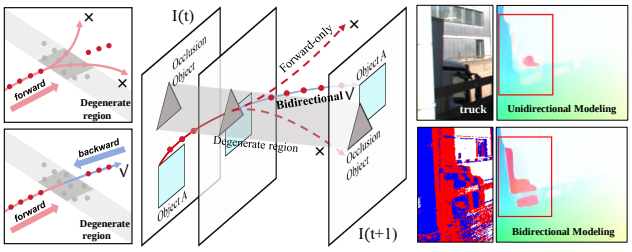

Figure 1. **Unidirectional vs. Bidirectional Temporal Modeling. Left:** Conceptual illustration of the unidirectional (top) and bidirectional (bottom) flow modeling approaches. **Middle:** Schematic illustration of the challenging occluded scene context. **Right:** Visualization of flow results in occluded scenes.

Beyond theoretical guarantees, bidirectional accumulation offers decisive practical advantages in handling "blind spots" such as occlusions and motion boundaries. As illustrated in Fig. 1, unidirectional models are fundamentally constrained to *extrapolate* motion into these regions—an inherently ill-posed problem given the loss of immediate visual evidence. In contrast, our bidirectional framework leverages future context to reframe this challenge as a well-posed *interpolation* task. This capability allows the BRU module to effectively "bridge" temporal gaps, recovering physically coherent trajectories even during temporary visibility loss. This structural advantage is empirically validated in Fig. 2, where our method resolves local ambiguities at sharp motion discontinuities and maintains trajectory integrity in complex occluded regions where unidirectional baselines typically falter.

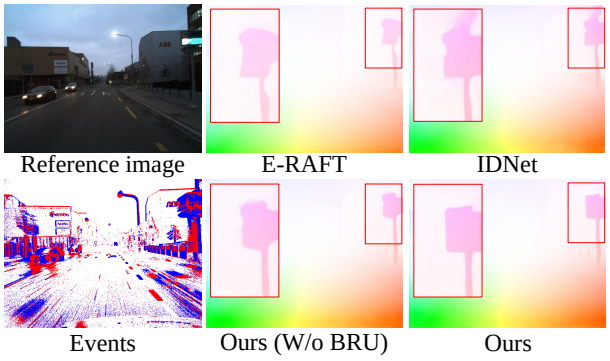

Figure 2. Qualitative comparisons in occluded and boundary-dominated regions on DSEC-Flow [1]. Red boxes indicate challenging areas.

B. Detailed Implementation Guidelines

To facilitate full reproducibility, we provide granular details regarding our training curriculum and architecture.

B.1. Network Hyperparameters

- **Bi^{2}ME:** The feature encoder outputs 64 channels for both the low-resolution $(1/8)$ and high-resolution $(1/4)$ branches.
- **Cost Volume:** We utilize a lookup radius of $r = 2$. This results in a local correlation volume of size $(2r+1)^2 \times T$ for each pixel.
- **Loss Weights:** For the supervised loss, we set $\gamma_1 = 0.25$ and $\gamma_2 = 0.75$.

B.2. Curriculum Schedule

We employ a dynamic weighting scheme. The supervised weight λ_{flow} linearly decays from $1.0 \rightarrow 0.1$ over the first 80% of epochs. Simultaneously, the self-supervised weights λ_{LSC} and λ_{TC} linearly ramp up from $0.0 \rightarrow 1.0$. This ensures the model first learns correct motion directions from sparse GT before refining trajectory continuity via STSC.

C. Extended Experimental Analysis

In this section, we provide a comprehensive analysis of our design choices, focusing on the trade-off between computational efficiency and performance, as well as the impact of data granularity.

C.1. Efficiency and Convergence Analysis

We critically evaluate the trade-off between computational overhead and estimation accuracy using the data in Table 1 and Table 2.

Computational Cost vs. Accuracy. Our bidirectional, dual-scale architecture prioritizes reconstruction fidelity. As shown in Table 1, although our inference latency is higher than lightweight baselines, this computational investment yields a decisive return in performance. Our method achieves the lowest End-Point Error (EPE) of **0.663**, significantly outperforming faster methods like BFlow (0.750) and TMA (0.743). Crucially, our model remains highly parameter-efficient, requiring fewer parameters (4.4M) than leading transformer-based or recurrent baselines like TMA (6.9M) and E-RAFT (5.3M). This demonstrates that our approach effectively allocates capacity to where it matters most—accuracy and structural consistency—rather than simply inflating model size.

Table 1. Performance vs. Efficiency on DSEC-Flow [1]. While incurring higher latency than lightweight baselines, our method achieves the best accuracy (lowest EPE) with a highly competitive parameter count.

Method	EPE ↓	Time [ms]	Params [M]	Memory
E-RAFT [2]	0.788	47	5.3	660 MB
TMA [5]	0.743	28	6.9	1.10 GB
BFlow [3]	0.750	64	5.3	1.41 GB
IDNet [7]	0.719	101	2.5	645 MB
Ours	**0.663**	128	4.4	701 MB

Rapid Convergence. Complementing this analysis, Table 2 highlights a key efficiency advantage of our design: rapid convergence. Unlike standard recurrent architectures that often require long update chains (e.g., 12 iterations) to refine flow estimates, our method saturates in performance at just **4 iterations**. The bidirectional context provides a robust global initialization that obviates the need for extensive iterative refinement, thereby effectively mitigating the per-iteration computational cost.

Table 2. Impact of the number of BRU iterations. Our model achieves optimal performance with significantly fewer iterations than typical RAFT-like approaches.

Iters	EPE	3PE	2PE	1PE	AE
1	0.773	2.36	4.08	12.35	2.87
2	0.687	1.84	2.97	8.76	2.64
3	0.667	1.64	2.78	7.97	2.55
4	**0.663**	**1.60**	**2.67**	**7.94**	**2.53**
5	0.671	1.66	2.71	8.01	2.52

C.2. Ablation on System Design

Effect of Event-Splitting Granularity. We investigate the impact of the temporal resolution used to discretize the input event stream. As shown in Table 3, increasing the number of bins B from 1 to 15 yields a monotonic improvement across all metrics. Using a single bin ($B = 1$) collapses all temporal information, leading to significant aliasing. In contrast, a higher granularity ($B = 15$) allows the model to effectively capture the microsecond-level evolution of nonlinear motion dynamics. We therefore adopt $B = 15$ as the optimal setting.

Table 3. Ablation on the number of event splits (B). Finer granularity yields better motion resolution.

Event Split (B)	EPE	3PE	2PE	1PE	AE
1	0.714	2.08	3.45	9.84	2.81
3	0.705	1.90	3.22	9.14	2.69
5	0.678	1.68	2.81	8.22	2.60
15	**0.663**	**1.60**	**2.67**	**7.94**	**2.53**

Component Analysis of Bi^{2}ME. Table 4 isolates the contributions of the key components within our Bidirectional Bi-scale Motion Encoder. Removing the Motion-Aware Difference (MADiff) increases EPE to 0.679, confirming its necessity for capturing high-frequency local details. Furthermore, removing the Global Correlation volume results in a degradation to 0.688 EPE, demonstrating that long-range matching priors are indispensable for anchoring motion estimation in the presence of large displacements.

Table 4. Ablation of Bi^{2}ME components.

Method	EPE	3PE	2PE	1PE	AE
w/o MADiff	0.679	1.69	2.81	8.22	2.58
w/o Correlation	0.688	1.77	3.02	8.54	2.62
Ours	**0.663**	**1.60**	**2.67**	**7.94**	**2.53**

Search Radius. We analyze the local search radius used in the correlation lookup. As reported in Table 5, a radius of 2 provides the best performance. Smaller radii ($r = 1$) limit the receptive field, failing to capture larger motions (EPE 0.672), while larger radii ($r = 4$) introduce matching ambiguity from distant, irrelevant regions (EPE 0.674). A radius of 2 offers the optimal trade-off between context capture and matching precision.

Table 5. Ablation on cost volume search radius.

Radius	EPE	3PE	2PE	1PE	AE
1	0.672	1.71	2.73	8.23	2.55
2	**0.663**	**1.60**	**2.67**	**7.94**	**2.53**
3	0.667	1.63	2.76	8.09	2.59
4	0.674	1.74	2.81	8.13	2.54

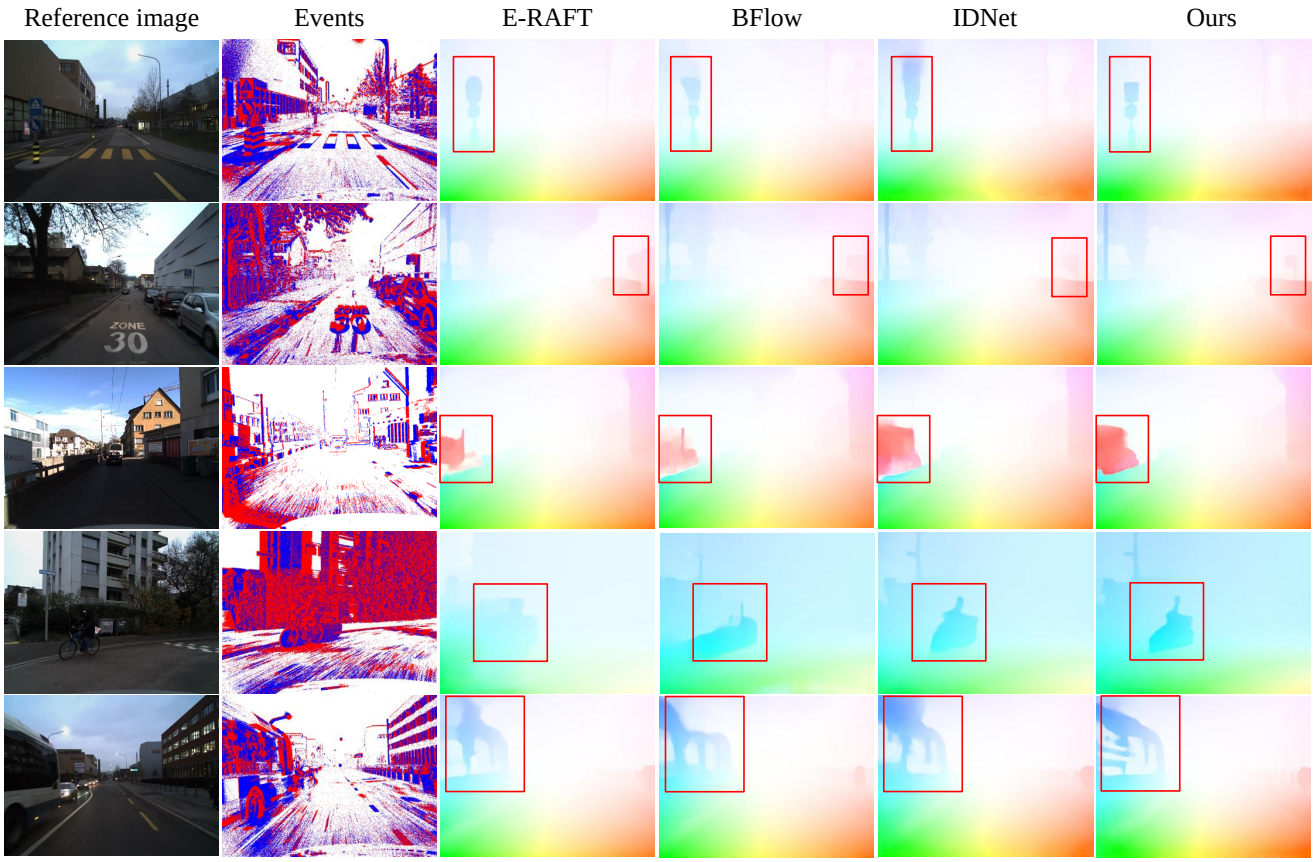

Figure 3. **Additional Qualitative Results on DSEC-Flow.** Comparing our method against E-RAFT, BFlow, and IDNet across diverse scenes. Our method demonstrates superior boundary preservation and reduced artifacts in dynamic regions.

D. Additional Qualitative Visualizations

To further substantiate the quantitative gains reported in the main text, we present an extended set of qualitative comparisons on the DSEC-Flow benchmark [1].

D.1. Performance in Diverse Scenarios

In Fig. 3, we extend our visual analysis to a broader spectrum of challenging scenarios, ranging from high-speed ego-motion to scenes populated with intricate structural details and small, independently moving objects. We benchmark our approach against three representative state-of-the-art methods: the correlation-based E-RAFT [2], the lightweight IDNet [7], and the continuous-time specialist BFlow [3]. As observed, baseline methods frequently struggle to maintain clear separation between foreground and background motion, often exhibiting characteristic failure modes such as over-smoothed boundaries or trajectory deviations in texture-less regions. In contrast, our method demonstrates superior structural fidelity and temporal coherence. By effectively integrating bi-scale spatial features with bidirectional temporal context, our model consistently recovers sharper motion boundaries and preserves the integrity of fine-grained structures, effectively mitigating the motion bleeding and ghosting artifacts prevalent in prior approaches.

D.2. Flow Error Distribution

To provide a granular assessment of estimation accuracy, we visualize per-pixel End-Point Error (EPE) heatmaps in Fig. 4, where darker intensities denote higher errors. Relative to the recent state-of-the-art IDNet [7], our method exhibits significantly lighter heatmaps across the image domain. The reduction in error is particularly pronounced along motion boundaries and object contours, confirming that our architectural designs—specifically the bi-scale encoder and bidirectional refinement—translate into genuine improvements in fine-grained motion preservation.

E. Limitations and Future Work

While our proposed framework sets a new standard for accuracy and robustness in event-based continuous-time optical flow, we acknowledge certain limitations that open av-

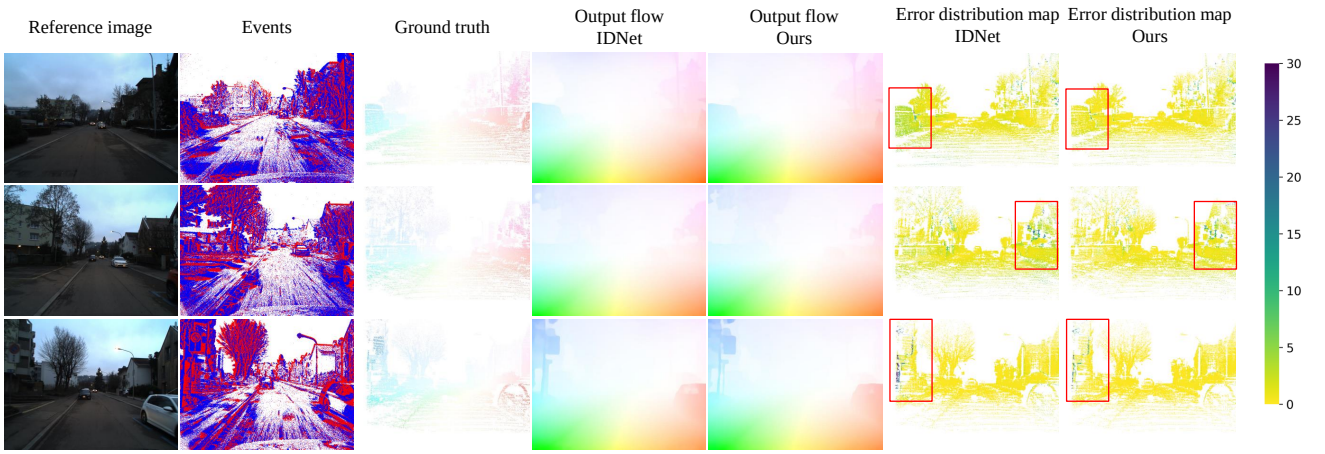

Figure 4. **Flow Error Maps.** Heatmaps encoding the End-Point Error (EPE), with darker colors representing higher errors. Our method shows significantly reduced error magnitude compared to IDNet, especially along motion boundaries.

enues for future research.

Computational Overhead. Our bidirectional, dual-scale design prioritizes reconstruction fidelity and physical consistency, which naturally incurs a higher computational cost compared to lightweight unidirectional baselines. Although our method remains parameter-efficient, the increased latency may currently constrain deployment on strictly power-limited edge devices requiring ultra-high-frequency processing. Future work will explore architectural distillation and sparse optimization to reduce runtime while preserving the structural integrity provided by our bidirectional modeling.

Extension to Longer-Term Dynamics. Currently, our STSC framework enforces consistency within a local spatio-temporal window. However, complex dynamics often evolve over timescales exceeding a single window's duration. Future research could extend our approach by enforcing **inter-window trajectory continuity**—for instance, by constraining the Bézier control points across consecutive windows to ensure smooth transitions or by incorporating long-term recurrent memory. This would enable the model to maintain coherent tracking through prolonged occlusions and recover consistent motion paths over extended sequences, moving towards a unified global representation of event dynamics.

References

[1] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5633–5643, 2019. 2, 3, 4

[2] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 197–206. IEEE, 2021. 1, 3, 4

[3] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4736–4746, 2024. 3, 4

[4] Daikun Liu, Lei Cheng, Teng Wang, and Changyin Sun. Edcflow: Exploring temporally dense difference maps for event-based optical flow estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1984–1993, 2025.

[5] Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhijun Li, Alois Knoll, and Changjun Jiang. Tma: Temporal motion aggregation for event-based optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9685–9694, 2023. 3

[6] Xinglong Luo, Kunming Luo, Ao Luo, Zhengning Wang, Ping Tan, and Shuaicheng Liu. Learning optical flow from event camera with rendered dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9847–9857, 2023.

[7] Yilun Wu, Paredes-Vallés, Federico, De Croon, and Guido CHE. Lightweight event-based optical flow estimation via iterative deblurring. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14708–14715. IEEE, 2024. 1, 3, 4